(19) Republik Österreich Patentamt

(11) Nummer: E 72 083 B

Best Available Copy

Übersetzung der europäischen

(12) # PATENTSCHRIFT

Veröffentlichungsnummer: 0 225 729 B1

(21) Anmeldenummer: 86308732

(51) Int.Cl.[5]: H04N 7/137

(22) Anmeldetag: 10.11.1986

(45) Ausgabetag: 25. 8.1992

(54) BILDKODIERUNG UND SYNTHESE.

(30) Priorität:

14.11.1985 GB 8528143

(43) Veröffentlichungstag der Anmeldung:

16. 6.1987, Patentblatt 87/25

(45) Bekanntmachung des Hinweises auf die Patenterteilung:

22. 1.1992, Patentblatt 92/04

(84) Benannte Vertragsstaaten:

AT BE CH DE ES FR GB GR IT LI LU NL SE

(56) Entgegenhaltungen:

BELL LABORATORIES RECORD, VOL. 48, NO. 4, APRIL 1970, PAGES 110-115, MURRY HILL, US; F.W: MOUNTS: ''CONDITIONAL REPLENISHMENT: A PROMISING TECHNIQUE FOR VIDEO TRANSMISSION''

(73) Patentinhaber:

BRITISH TELECOMMUNICATIONS PUBLIC LIMITED COMPANY
81 NEWGATE STREET
LONDON EC1A 7AJ (GB).

(72) Erfinder:

WELSH, WILLIAM JOHN
47, FOUNTAINS ROAD
IPSWICH SUFFOLK, IP2 9EF (GB).

FENN, BRIAN ALAN
43, CATHERINE ROAD
WOODBRIDGE SUFFOLK IP12 4JP (GB).

CHALLENER, PAUL
10, FREEMAN AVENUE
HENLEY IPSWICH, SUFFOLK, IP6 0RZ (GB).

Anmerkung:

Innerhalb von neun Monaten nach der Bekanntmachung des Hinweises auf die Erteilung des europäischen Patents im Europäischen Patentblatt kann jeder beim Europäischen Patentamt gegen das erteilte europäische Patent Einspruch einlegen. Der Einspruch ist schriftlich einzureichen und zu begründen. Er gilt erst als eingelegt, wenn die Einspruchsgebühr entrichtet worden ist (Art. 99(1) Europäisches Patentübereinkommen).

Die Übersetzung ist gemäß § 5 PatVEG vom Patentinhaber eingereicht worden. Sie wurde vom Österreichischen Patentamt nicht geprüft!

E 72 083 B

86 308 732.6
British Telecommunications plc

20. Januar 1992
B 14927 DE AL/Kh/sb

Die vorliegende Erfindung bezieht sich auf das Codieren von sich bewegenden Bildern, auf welchen ein menschliches Gesicht dargestellt ist. Sie befaßt sich mit dem Erzielen von geringen Übertragungsraten, indem sie sich auf Bewegungen konzentriert, die mit der Sprache zusammenhängen. Die Erfindung gestattet auch die Synthese derartiger Bilder, um wirkliche oder synthetische Sprache zu begleiten.

Es wurde schon vorgeschlagen (siehe Bell Laboratories Record, Band 48, Nr. 4, April 1970, Seiten 110 bis 115, Murry Hill, US; F.W. Mounts: "Conditional Replenishment: A Promising Technique For Video Transmission"), die benötigte Übertragungsrate für ein sich bewegendes Bild zu verringern, indem man aufeinanderfolgende Datenblöcke des Bildes vergleicht und Daten nur hinsichtlich derjenigen Teile des Datenblockes überträgt, welche sich seit dem vorhergehenden Datenblock geändert haben. Die vorliegende Erfindung zielt darauf ab, die Kenntnis auszunutzen, daß beim Übertragen eines Bildes eines Gesichts der Hauptinformationsgehalt in den Bewegungen des Mundes liegt.

Gemäß eines ersten Aspektes der Erfindung ist ein Gerät zum Codieren eines sich bewegenden Bildes bereitgestellt einschließlich eines menschlichen Gesichts, wobei das Gerät aufweist:

eine Einrichtung zum Empfangen von Videoeingabedaten;

eine Einrichtung zur Ausgabe von Daten, welche einen Datenblock des Bildes darstellen;

eine Identifikationseinrichtung, welche angeordnet ist, um im Betrieb für jeden Datenblock des Bildes denjenigen Teil der Eingabedaten zu identifizieren, welche dem Mund des dargestellten Gesichts entsprechen, und

a) um in einer ersten Betriebsphase die Munddatenteile jedes Datenblocks mit denen anderer Datenblöcke zu vergleichen, um einen repräsentativen Satz von Munddatenteilen auszuwählen, den repräsentativen Satz zu speichern und diesen Satz auszugeben;

b) um in einer zweiten Phase die Munddatenteile jedes Datenblocks mit denen des gespeicherten Satzes zu vergleichen und zum Erzeugen eines auszugebenden Codeworts, welches anzeigt, welchem Element des Satzes die Munddatenteile des Datenblockes am meisten ähneln.

Man wird zu schätzen wissen, daß dieses Vorgehen sich zuvor bekannten Wissens hinsichtlich der Natur des Bildes bedient, indem speziell der Mund des dargestellten Gesichts identifiziert wird, und nutzt weiterhin die Tatsache aus, daß der Mund angemessen dargestellt werden kann durch einen ausgewählten repräsentativen Satz von Munddatenteilen.

Gemäß eines zweiten Aspekts der Erfindung ist ein Sprachsynthetisator bereitgestellt mit einer Einrichtung für die Synthese eines sich bewegenden Bildes einschließlich eines menschlichen Gesichts, wobei der Sprachsynthetisator aufweist:

a) eine Einrichtung zum Speichern und Ausgeben des Bildes eines Gesichts;

b) eine Eingabe zum Speichern und Ausgeben eines Satzes von Munddatenblöcken (Fig. 3), die jeweils dem Mundgebiet des Gesichts entsprechen und eine jeweilige unterschiedliche Mundform darstellen;

c) eine Eingabe zum Empfangen von Codes, welche Worte oder Teile von zu sprechenden Worten identifizieren;

d) eine Sprachsyntheseeinrichtung, welche auf den an der Eingabe empfangenen Code anspricht, um dazu entsprechende Worte oder Teile von Worten zu synthetisieren;

e) eine Einrichtung, die eine Tabelle speichert, welche derartige Codes mit Codeworten in Verbindung setzt, welche die Munddatenblöcke oder Sequenzen derartiger Codeworte identifiziert; und

f) eine Steuereinrichtung, welche auf die an der Eingabe empfangenen Codes anspricht, um das entsprechende Codewort oder die Codewortsequenz von der Tabelle auszuwählen und sie synchron mit der Synthese des entsprechenden Wortes oder Teils eines Wortes durch die Sprachsyntheseeinrichtung auszugeben.

Gemäß eines dritten Aspektes der Erfindung ist ein Gerät bereitgestellt zur Synthese eines sich bewegenden Bildes, wobei das Gerät aufweist:

a) eine Einrichtung zum Speichern und Ausgeben des Bildes eines Gesichts;

b) eine Einrichtung zum Speichern und Ausgeben eines Satzes von Munddatenblöcken, die jeweils dem Mundgebiet des Gesichts entsprechen und eine jeweilige unterschiedliche Mundform darstellen;

c) eine Audio-Eingabe zum Empfangen von Sprachsignalen und eine Frequenzanalyse-Einrichtung, welche auf derartige Signale anspricht zum Erzeugen von Sequenzen spektraler Parameter;

d) eine Einrichtung, die eine Tabelle speichert, welche spektrale Parametersequenzen mit Codeworten in Beziehung setzt, wobei Munddatenblöcke oder Sequenzen davon identifiziert werden;

e) eine Steuereinrichtung, die auf die spektralen Parameter anspricht, um für eine Ausgabe die entsprechenden Codeworte oder Codewortsequenzen von der Tabelle auszuwählen.

Einige Ausführungsbeispiele der Erfindung werden nun beispielhaft beschrieben unter Bezugnahme auf die begleitende Zeichnung.

Fig. 1 ist ein Blockdiagramm eines Bildübertragungssystems einschließlich eines Codierers und Empfängers gemäß den Ausführungsbeispielen der Erfindung;

Fig. 2 zeigt ein zu übertragendes Bild;

Fig. 3 zeigt einen Satz von Mundformen;

Fig. 4, 5 & 6 zeigen Maskierungsfenster, welche bei der Gesichts-, Augen- und Mund-Identifikation verwendet werden;

Fig. 7 ist ein Histogramm, welches durch Verwendung der Maske von Fig. 6 erhalten worden ist;

Fig. 8 & 9 zeigen binäre Bilder des Mundgebietes eines Bildes;

Fig.10 & 11 sind Grundriß- und Aufrißansichten eines Kopfes, um die Effekte von Änderungen der Orientierung darzustellen;

Fig. 12 zeigt ein Gerät zur Sprachanalyse;

Fig. 13 ist ein Blockdiagramm eines Empfängers, der die Erfindung verkörpert.

Fig. 1 zeigt ein Bildübertragungssystem mit einem Sender 1, einer Übertragungsverbindung 2 und einem Empfänger 3. Die verwendeten Techniken sind gleichermaßen anwendbar für ein Aufzeichnen, und die Übertragungsverbindung 2 könnte daher durch ein Bandaufzeichnungsgerät oder eine andere Einrichtung, wie z.B. einem Halbleiterspeicher, ersetzt werden.

Der Sender 1 empfängt ein Eingabevideosignal von einer Quelle, wie z.B. einer Kamera.

Das sich bewegende zu übertragende Bild ist das Gesicht 5 (Fig. 2) eines Sprechers, dessen Sprache auch über die Verbindung 2 zu dem Empfänger übertragen wird. Während gewöhnlichem Sprechen gibt es verhältnismäßig wenig Änderung im größten Teil der Gesichtsfläche - d.h. dem nicht dem Mundgebiet angehörigen Teil, der durch den Kasten 6 in

Fig. 2 angedeutet ist. Daher wird nur ein Bild des Gesichts übertragen. Weiterhin findet man, daß Änderungen in den Mundpositionen während dem Sprechen realistisch dargestellt werden können unter Verwendung einer relativ kleinen Anzahl verschiedener Mundpositionen, welche als typisch ausgewählt werden. Dadurch wird ein Code-Buch von Mundpositionen erzeugt, und, wenn dies einmal zu dem Empfänger übertragen worden ist, ist die einzige weitere Information, welche gesendet werden muß, eine Sequenz von Codeworten, welche die aufeinanderfolgenden darzustellenden Mundpositionen identifizieren.

Das beschriebene System ist ein auf Kenntnis basierendes System - d.h. von dem Empfänger wird nach einer "Lern"-Phase angenommen, daß er das Gesicht des Sprechers und den Satz von Mundpositionen "kennt". Der Betrieb des Empfängers ist unumständlich und involviert in der Lernphase eine Eingabe des Gesichtsbildes in einen Datenblockspeicher (von welchem ein Ausgabevideosignal erzeugt wird durch wiederholtes Auslesen) und eine Eingabe des Satzes von Mundpositionen in einen weiteren "Mund"-Speicher und in der Übertragungsphase eine Verwendung von jedem empfangenen Codewort, um die angemessenen Mundbilddaten wiederzugewinnen und das entsprechende Gebiet des Bildspeichers zu überschreiben.

Der Senderbetrieb ist notwendigerweise komplexer, und hierbei benötigt die Lernphase eine Übungssequenz von dem Sprecher, wie folgt:

1) der erste Datenblock wird gespeichert und in passender Weise codiert (z.B. unter Verwendung konventioneller Redundanzverringerungstechniken) zu dem Empfänger übertragen.

2) Das gespeicherte Bild wird analysiert, um (a) den Kopf des Sprechers zu identifizieren (so daß der Kopf in zukünftigen Datenblöcken trotz Kopfbewegungen abgeglichen werden kann), und um (b) den Mund zu identifizieren - d.h. definieren des Kastens 6, welcher in Fig. 2 gezeigt ist. Die Kastenkoordinaten (und Dimensionen, falls nicht festgelegt) werden zu dem Empfänger übertragen.

3) Aufeinanderfolgende Datenblöcke der Übungssequenz werden analysiert, um den Mund abzugleichen und dadurch die momentane Position des Kastens 6 zu definieren und den Inhalt des Kastens (des "Mundbildes") mit dem ersten und jedem zuvor ausgewählten Bild zu vergleichen, um einen Satz ausgewählter Mundbilder aufzubauen. Dieser Satz von (in Fig. 3 dargestellten) Bildern wird bei dem Sender gespeichert und zu dem Empfänger übertragen.

Die Übertragungsphase benötigt dann:

4) ein Analysieren aufeinanderfolgender Datenblöcke (wie in (3) oben), um die Position des Kastens 6 zu identifizieren;

5) ein Vergleichen des Inhalts des Kastens in dem momentanen Datenblock mit den gespeicherten Mundbildern, um dasjenige des Satzes zu identifizieren, welches ihm am nächsten ist; woraufhin das entsprechende Codewort übertragen wird.

Nimmt man eine Datenblockrate von 25 pro Sekunde an und ein "Codebuch" von 24 Mundformen (d.h. einen 5-Bit Code), würde die benötigte Datenrate während der Übertragungsphase 125 Bits/s betragen.

Die unter Verwendung des grundlegenden beschriebenen Systems erzielte Empfängeranzeige stellt sich als allgemein zufriedenstellend heraus, aber ist etwas unnatürlich, hauptsächlich weil (a) der Kopf feststehend erscheint und (b) die Augen unverändert bleiben (genau gesagt, der Sprecher scheint niemals zu zwinkern). Dem ersten dieser Probleme kann abgeholfen werden, indem man eine zufällige Kopfbewegung bei dem Empfänger einführt; oder durch Abgleichen der Kopfposition bei dem Sender und Übertragen von angemessenen Koordinaten zu dem Empfänger. Die Augen könnten übertragen werden unter Verwendung der gleichen Prinzipien wie sie für den Mund angewendet werden; obwohl hier die Größe des "Codebuches" viel kleiner sein kann. Ähnliche Bemerkungen treffen für das Kinn und Gesichtszüge zu.

Die Implementierung der oben aufgezählten Senderschritte wird nun in etwas detaillierterer Weise betrachtet, unter Annahme eines monochromen Quellbildes von 128x128 pel Auflösung eines Bildes von Kopf und Schultern. Das erste Problem liegt im Erkennen von Gesichtsmerkmalen und deren Festlegung auf dem Gesicht. Andere Probleme bestehen im Bestimmen der Orientierung des Kopfes und der sich ändernden Form des Mundes sowie der Bewegung der Augen. Das von Nagao vorgeschlagene Verfahren (M. Nagao - "Picture Recognition And Data Structure", Graphic Languages - E.D. Nake & Rosenfield, 1972) wird vorgeschlagen.

Nagaos Verfahren involviert die Erzeugung einer binären Darstellung des Bildes mit einem Kantendetektor. Dieses binäre Bild wird dann analysiert, indem ein Fenster an ihm herunterbewegt wird und die Kantenpixel in jeder Spalte des Fensters summiert werden. Die Ausgabe des Fensters ist der Satz Zahlen, in welchem große Zahlen starke vertikale Kanten darstellen. Daraus können Merkmale, wie z.B. die Spitze und die

Seiten des Kopfes, gefolgt von den Augen, der Nase und dem Mund anfänglich erkannt werden.

Der Algorithmus fährt fort und bestimmt den Umriß des Kiefers und arbeitet sich dann an dem Gesicht hoch, um die Positionen von Nase, Augen und Seiten des Gesichts genauer festzulegen. Ein in den Algorithmus eingebauter Rückkoppelungsprozess gestattet eine Wiederholung von Teilen der Suche, falls ein Fehler erfaßt wird. Auf diese Art und Weise wird die Erfolgsrate stark verbessert.

Ein Programm wurde unter Verwendung des Algorithmus von Nagao geschrieben, welches Rechtecke um die als Augen und Mund identifizierten Merkmale zeichnet. Es folgen Details dieses Programms:

Ein Laplace-Operator wird angewandt zusammen mit einem Schwellwert, um ein binäres Bild gleicher Auflösung zu geben. Kantenpixel werden schwarz, andere weiß.

Ein Fenster der Dimension 128 pel x 8 Zeilen ist an dem Oberteil des binären Bildes positioniert. Die schwarzen Bildelemente bzw. Pixels bzw. pels in jeder Spalte werden summiert, und das Ergebnis wird als eine Eingabe in einer Anordnung von 128 x 32 Elementen (Array 1) gespeichert. Das Fenster wird um vier Zeilen jedesmal bildabwärts bewegt und der Prozess wiederholt. Das Fenster wird insgesamt 32 mal neu positioniert, und die Anordnung von 128 x 32 Elementen wird gefüllt (Fig. 4).

Eine Suche wird durchgeführt durch die Reihen von Array 1, und zwar beginnend von dem Oberteil des Bildes, um die Seiten des Kopfes zu

lokalisieren. Da dies starke vertikale Kanten sind, werden sie durch hohe Werte in Array 1 identifiziert.

Die erste von der linken Seite des Bildes lokalisierte Kante wird aufgenommen, und ähnlich geht man für die rechte Seite vor. Der Abstand zwischen diesen Punkten wird gemessen (Kopfbreite), und falls dieser Abstand ein Kriterium übersteigt, wird eine Suche nach Aktivität zwischen diesen beiden Punkten durchgeführt, welche die Augen anzeigen kann.

Die Augen werden gefunden, indem man eine eindimensionale Maske verwendet, wie in Fig. 5 gezeigt, welche zwei Schlitze hat, die den durch einen Spalt für die Nase getrennten Augen entsprechen. Die Breite der Schlitze und ihr Abstand wird so ausgewählt, daß sie proportional zu der gemessenen Kopfbreite ist. Die Maske wird an einer Reihe innerhalb der Kopffläche entlang bewegt. Die Zahlen in Array 1, welche innerhalb der Augenschlitze fallen, werden summiert, und von diesem Ergebnis werden die Zahlen in dem Nasenschlitz subtrahiert. Das Endergebnis ist ein empfindlicher Indikator von Aktivität aufgrund der Augen.

Der Maximalwert entlang einer Reihe wird aufgezeichnet zusammen mit der Position der Maske, wenn dieses Maximum gefunden wird. Die Maske wird dann nach unten bewegt zu der nächsten Reihe und der Prozess wiederholt.

Aus dem Satz von Maximalwerten wird das Gesamtmaximum gefunden. Die Position dieses Maximums wird als Angabe der vertikalen Position der Augen betrachtet. Verwendet man die horizontale Position der Maske, wenn dieses Maximum gefunden wurde, können wir den Mittelpunkt des Gesichts abschätzen.

Danach wird ein 15 Pixel-weites Fenster (Fig. 6) auf das binäre Bild angewendet. Es erstreckt sich von einer Position genau unterhalb der Augen zu dem Unterteil des Bildes und ist auf der Mitte des Gesichts zentriert.

Die schwarzen pels in jeder Reihe des Bildes werden summiert, und die Werte werden in eine eindimensionale Anordnung (Array 2) eingegeben. Falls diese Anordnung als ein Histogramm angezeigt wird, werden Merkmale, wie der Unterteil der Nase, der Mund und der Schatten unter der unteren Lippe, deutlich als Spitzen sichtbar (Fig. 7). Die Verteilung dieser Spitzen wird verwendet, um die Position des Mundes festzulegen.

Die Kastenposition wird als zentriert auf dem Zentrum des Gesichts bestimmt, wie oben definiert, und auf dem Zentrum des Mundes (Reihe 35 in Fig. 7). Für die gegebene Auflösung kann eine passende Kastengröße eine Breite von 40 x einer Höhe von 24 pels sein. Das nächste Stadium besteht darin, sicherzustellen, daß die Identifizierung des Mundes (Kastenposition) in dem ersten Datenblock und während der Lern- (und Übertragungs-)Phase konsistent ist - d.h., daß der Mund immer innerhalb des Kastens zentriert ist. Man findet, daß eine Anwendung des Algorithmus von Nagao auf jeden Datenblock einer Sequenz dagegen einen beachtlichen Fehler bei der Registrierung des Mundkastens von Datenblock zu Datenblock aufweist.

Eine Lösung dieses Problems wurde gefunden, indem man den Algorithmus nur auf den ersten Datenblock anwendet und dann den Mund Datenblock für Datenblock abgleicht. Dies wird erreicht, indem man den Mund in dem ersten Datenblock der binären Sequenz als eine Schablone benutzt und mit jedem der nachfolgenden Datenblöcke in dem obigen binären Bild auto-korreliert. Die Suche wird begonnen in dersel-

ben relativen Position in dem nächsten Datenblock, und die Maske wird jedesmal um ein Pixel bewegt bis ein lokales Maximum gefunden ist.

Das Verfahren wurde verwendet, um eine Sequenz zu erhalten unter Verwendung des korrekten Mundes, wobei aber der Rest des Gesichts von dem ersten Datenblock kopiert wird. Diese verarbeitete Sequenz ließ man laufen, und sie zeigte etwas Aufzeichnungsflackern, aber dieser Fehler betrug nur ungefähr ein Pixel, was das Beste ist, das man erreichen kann ohne Sub-Pixelinterpolation.

Typische binäre Bilder der Mundfläche (Mund offen und Mund geschlossen) sind in Fig. 8 und 9 gezeigt.

Nur ein kleiner Satz von Mündern von der gesamten möglichen Anzahl in der gesamten Sequenz kann in der Nachschlagetabelle aus offensichtlichen Gründen gespeichert werden. Dies benötigt, daß die Form eines Mundes erkannt wird und ob sie ähnlich einer Form ist, welche zuvor aufgetreten ist oder nicht. Neue Mundpositionen würden dann in der Tabelle gespeichert werden.

Die Ähnlichkeit oder der Unterschied eines Mundes zu vorhergehend auftretenden Mündern muß daher auf einem Quantisierungsprozess aufgebaut werden, um die Anzahl der Eintragungen in die Tabelle zu begrenzen.

Das Verfahren, wodurch dies erreicht wird, besteht wie im folgenden darin, daß die gesamte Verarbeitung auf Grauskalen-Mundbildern durchgeführt wird anstelle der obigen binären Version.

Das Mundbild von dem ersten Datenblock wird als die erste - zu Beginn die einzige - Eingabe in einer Nachschlagetabelle gespeichert. Das Mundbild von jedem Datenblock in der Übungssequenz wird dann verarbeitet, indem man es (a) mit jeder Eingabe in die Tabelle vergleicht durch Subtrahieren der individuellen pel-Werte und Summieren der absoluten Werte dieser Differenzen über dem Mundkastengebiert; indem man (b) die Summe mit einem Schwellenwert vergleicht und, falls der Schwellenwert überschritten wird, man dieses Mundbild als eine Neueingabe in die Tabelle eingibt.

Dieses besondere Verfahren zum Auffinden der Summe der absoluten Differenzen ist jedoch sehr empfänglich bzw. empfindlich für eine Bewegung. Zum Beispiel würden zwei identische Bilder, bei denen das zweite um nur ein Pixel nach links verschoben worden ist, einen sehr niedrigen Wert für die Summe erzeugen, wohingegen diese zwei Bilder als identisch angesehen werden sollten. Falls ein kleines Ausmaß an Bewegung innerhalb des gesamten Abgleichs gestattet wird, um zu versuchen, die Tatsache zu kompensieren, daß die Summe dramatisch abfällt, falls das Bild nur um ein Pixel verschoben worden ist, dann kann eine Verringerung der Größe der Nachschlagetabelle erzielt werden ohne einen entsprechenden Verlust von Mundformen. Dies kann durchgeführt werden, wenn bei jedem Datenblock der Mund in dem momentanen Datenblock dreimal mit jeder der Eintragungen in dem Codebuch verglichen wird - und zwar bei der momentanen Position, nach links um ein Pixel verschoben, und nach rechts um ein Pixel verschoben, und die minimale Summe in jedem Fall gefunden wird. Das Ergebnis, welches die kleinste minimale Summe erzeugt, zusammen mit dem Wert der Verschiebung in die X-Richtung wird aufgezeichnet. Diese Bewegung könnte natürlich sowohl in der X- als auch der Y-Richtung durchgeführt werden, aber

man stellte fest, daß die Mehrzahl der Bewegungen in der X-Richtung stattfinden.

Falls die gewünschte Tabellengröße überschritten wird, oder die Anzahl der Eintragungen, welche während der Übungssequenz erzielt werden, wesentlich kleiner ist als die Tabellengröße, dann wird der Schwellenwert entsprechend eingestellt und die Übungsphase wiederholt; um eine übermäßige Verzögerung zu vermeiden, können solche Bedingungen von der Erfassungsrate vorhergesagt werden.

Sobald die Tabelle einmal aufgebaut ist, kann die Übertragungsphase beginnen, in welcher jedes der aufeinanderfolgenden Mundbilder verglichen wird - wie in (a) oben beschrieben -, und zwar mit all denjenigen der gespeicherten Tabelle, und ein Codewort, welches die Eingabe identifiziert, die das geringste Summierungsergebnis ergab, wird dann übertragen.

Die dafür benötigte Berechnung ist umfangreich, kann aber verringert werden, falls man ein alternatives Suchverfahren übernimmt. Die einfachste Alternative wäre, daß man, anstatt alle Münder in der Nachschlagetabelle anzusehen und die minimale Summe zu finden, den ersten verwendet, der eine Summe hat, welche geringer ist als die Schwelle. Für sich allein wäre dies gewiß schneller, doch würde es wahrscheinlich darunter leiden, umfangreich verzerrt zu sein, falls die Reihenfolge, in der die Tabelle abgetastet wird, festgelegt wäre. Daher muß die Reihenfolge, in der die Tabelle abgetastet wird, variiert werden. Eine bevorzugte Variation benötigt, daß man die Reihenfolge aufzeichnet, in welcher die Munde von dem Codebuch erscheinen, eine Art Rangreihenfolge. Wenn z.B. der vorhergehende Datenblock den Mund 0 von der Tabelle verwendete, dann tastet man die Tabelle für den momentanen

Datenblock ab, wobei man mit der Eingabe beginnt, welche am meisten nach dem Mund 0 in der Vergangenheit aufgetreten ist, sagen wir Mund 5. Wenn die Summe der absoluten Differenzen zwischen dem momentanen Datenblock und Mund 5 weniger ist als die Schwelle, dann wird Mund 5 ausgewählt, um den momentanen Datenblock darzustellen. Falls sie größer ist als die Schwelle, bewegt man sich weiter zu dem nächsten Mund in dem Codebuch, welcher in dem Codebuch nach Mund 0 am zweithäufigsten erschienen ist, und so fort. Wenn ein Mund schließlich ausgewählt wird, wird die Aufzeichnung des ausgewählten Mundes auf den neuesten Stand gebracht, um die momentane Information zu beinhalten.

Wahlweise können Mundbilder mit einem niedrigsten Summierungsergebnis über einen eingestellten Wert als nicht in dem Satz vorhandene Formen erkannt werden und einen dynamischen Datenerneuerungsprozess einleiten, wobei ein zusätzliches Mundbild an die Tabelle angehängt wird und an den Empfänger während der Übertragungsphase gesendet wird. Bei den meisten Umständen wäre eine Übertragung eines "neuen" Mundes nicht schnell genug, um seine Verwendung für den ihn verursachenden Datenblock zu gestatten, sondern er stünde für zukünftige Erscheinungen dieser Form zur Verfügung.

In diesem Fall muß man aufpassen, daß der eingestellte Wert nicht zu tief ist, denn dies kann dazu führen, daß neue Münder während der ganzen Sequenz in die Nachschlagetabelle gebracht werden. Und dies ist nichts weiter als eine Bild-Teilabtastung, was offensichtlich ein vernünftiges Ergebnis erzeugen würde, was aber ein Codebuch benötigen würde, dessen Größe proportional ist zu der Länge der gerade verarbeiteten Sequenz.

Man kann zu dem eingestellten Wert durch wiederholtes Probieren gelangen. Offenbar wäre es wünschenswert, daß diese Schwelle automatisch ausgewertet werden könnte oder wenn man darauf völlig verzichten könnte. Die Summe aller absoluten Differenzen zwischen Datenblöcken ist immer ein positives Maß, und die Nachschlagetabelle stellt daher einen metrischen Raum dar. Man kann sich jeden Mund in der Nachschlagetabelle so vorstellen, als ob er in einem mehrdimensionalen metrischen Raum existierte, und jeder Datenblock in einer Sequenz liegt in einer Anhäufung um einen dieser Codebuch-Münder herum. Es gibt verschiedene Algorithmen, z.B. den Linde-Buzo-Gray-Algorithmus, die verwendet werden könnten, um den optimalen Satz von Mündern zu finden. Diese Algorithmen verwenden den Satz von Datenblöcken in der Sequenz als einen Übungssatz und verwenden langwierige Suchen, um den Fehler zu minimieren und den optimalen Satz zu finden. Dem ist vorzuziehen, einen "repräsentativen" Satz von Mündern zu finden, welche sub-optimal sind, aber welche schneller gefunden werden können als der optimale Satz. Um dies zu tun, ist es notwendig, die Anzahl der zu verwendenden Münder anzugeben, und dann die benötigte Anzahl von Mündern von der Übungssequenz auszuwählen. Die Nachschlagetabelle kann während der Übertragungsphase noch auf den neuesten Stand gebracht werden, wobei der gleiche Algorithmus verwendet wird, wie zum Üben, aber die gesamte Anzahl der Münder in der Tabelle wird konstant bleiben.

Die Auswahl von Mündern folgt einer Grundregel - falls der minimale Abstand (Abstand kann verwendet werden, da es ein metrischer Raum ist) zwischen dem momentanen Datenblock und einem der Münder in der Tabelle größer ist als der minimale Abstand zwischen diesem Mund in der Tabelle und irgendeinem anderen Mund in der Tabelle, dann sollte der momentane Mund in der Tabelle aufgenommen werden. Falls

er geringer ist, dann sollte dieser Mund einfach durch den nächstgelegenen Mund der Tabelle dargestellt werden. Wenn ein neuer Mund in die Tabelle während einer Übertragungsphase aufgenommen werden soll, dann wird der zu entfernende Mund nach der folgenden Regel ausgewählt - Finden des Paars von Munden in der Nachschlagetabelle, welche am nächsten zueinander sind und Wegwerfen eines der beiden, vorzugsweise desjenigen, der am nächsten zu dem neuen Mund liegt.

Wenn ein neuer Mund in die Tabelle eingegeben wird, dann hat er klarerweise keine Vorgeschichte, womit er die anderen Münder in das Codebuch einordnen könnte - jeder wird nach diesem neuen Mund niemals aufgetreten sein. Wenn der nächste Datenblock in der Sequenz angetroffen wird, würde die Nachschlagetabelle der Reihenfolge nach abgetastet werden, wobei die neue Eingabe zuletzt erreicht wird. Diese neue Eingabe ist jedoch die wahrscheinlichste Auswahl, da Münder dazu neigen, in Anhäufungen zu erscheinen, insbesondere wenn gerade ein neuer Mund erzeugt worden ist. So paßt man die Reihenfolge an, so daß der neue Mund als erstes abgetastet wird.

Das oben beschriebene Übertragungssystem kann in einem Bildtelefonsystem verwendet werden, welches eine Standardtelefonverbindung verwendet; um der Lernphase gerecht zu werden, würde das Bild nicht unmittelbar am Empfänger erscheinen. Der anfänglichen Verschiebung folgend - vielleicht 15 Sekunden, wenn man eine nicht-digitale Übertragung des Gesichts annimmt -, würde das bewegte Bild übertragen und im Echtzeitverfahren dargestellt werden.

Eine feste Mundüberlappung kann auf einem Gesicht verwendet werden, welches von der nach vorne zeigenden Position abweicht, falls die Differenz nicht zu groß ist. Es ist auch klar, daß zum Anzeigen allgemeiner

Kopfbewegungen, wie z.B. Nicken oder Schütteln, man das Gesicht so darstellen muß, wie man es von einer Anzahl verschiedener Winkel sieht. Ein dargestellter Kopf ist unüberzeugend, wenn nicht eine allgemeine Bewegung vorliegt, sei es auch nur eine wahllose Bewegung.

In einem System wie dem beschriebenen müßten verschiedene Ansichten des Gesichtes übertragen und bei dem Empfänger gespeichert werden. Falls ein kompletter Satz von Daten für jede unterschiedliche Gesichtsposition gesendet würde, würde dies eine exzessive Kanal- und Speicherkapazität erfordern. Ein möglicher Weg um dieses Problem herum ist in Fig. 10 gezeigt.

Das Erscheinen des Gesichtes in der frontalen Position wird dargestellt durch die Projektion (x1 - x5) in der Ebene P. Wenn der Kopf leicht auf eine Seite gedreht wird, wird sein Erscheinen für den Beobachter nun dargestellt durch (x1' - x5') in Ebene P'. Wenn die Beleuchtung des Gesichtes ziemlich isotropisch ist, dann sollte eine zweidimensionale Transformation (x1 - x5) eine gute Annäherung sein an (x1' - x5').

Die wichtigen Unterschiede würden an den Seiten des Kopfes auftreten, wo neue Flächen enthüllt oder verdeckt werden, in ähnlicher Weise bei der Nase. Damit kann durch Übertragung eines Codes, der die Änderung der Orientierung des Kopfes als auch einen kleinen Satz von Unterschieden gibt, der ganze Kopf rekonstruiert werden. Die Differenzen für jede Kopfposition können gespeichert werden und in der Zukunft verwendet werden, wenn die gleiche Position identifiziert wird.

Das Konzept, Pseudorotationen zu erzeugen durch 2-D Transformationen wird dargestellt mit Bezug auf das "Gesicht"-Bild von Fig. 11.

Simulieren des Effektes von vertikal-Achsenrotation in einer Richtung, so daß die Nase sich um eine Verschiebung S von links nach rechts (wie betrachtet) bewegt:

(1) Punkte links von (X1 - X1') bewegen sich nicht.
(2) Punkte auf der Linie (X2 - X2') bewegen sich nach rechts mit den Verschiebungen S/2. (Das Gebiet (X1, X1', X2, X2') wird entsprechend gestreckt.)
(3) Punkte auf der Linie (X3 - X3') bewegen sich nach rechts mit Verschiebung S. (Das Gebiet (X2, X2', X3, X3') wird gestreckt.)
(4) Punkte auf der Linie (X4 - X4') bewegen sich nach rechts um eine Verschiebung S. (Das Gebiet (X3, X3', X4, X4') wird nach rechts verschoben.)
(5) Punkte auf der Linie (X5 - X5') bewegen sich nach rechts; Verschiebung S/2. (Das Gebiet (X4, X4', X5, X5') wird geschrumpft.)
(6) Punkte rechts von der Linie (X6 - X6') bewegen sich nicht. (Das Gebiet (X5, X5, X6, X6') wird geschrumpft.)

Zweidimensionale graphische Transformationen könnten verwendet werden in einem System für eine Standardvideo-Konferenz-Anwendung. In diesem System würden menschliche Gegenstände erkannt werden und von sich nicht bewegenden Vordergrund- und Hintergrundgegenständen isoliert werden. Der Vordergrund und der Hintergrund würden in Speichern bei verschiedenen hierarchischen Niveaus abgespeichert werden je nach dem, ob sie in der Lage wären, sich bewegende Gegenstände zu verdecken. Sich relativ wenig verändernde bewegende Körper, wie z.B. Rümpfe, würden bei einem anderen Niveau gespeichert werden als sich schneller ändernde Teile, wie z.B. die Arme und der Kopf.

Das Betriebsprinzip des Systems würde von dem Übertragungsende verlangen, daß es die Bewegung verschiedener segmentierter Teile identifiziert und Bewegungsvektoren entsprechend sendet. Diese würden von dem Empfänger verwendet, um eine Vorhersage für jeden Teil in dem nächsten Datenblock zu bilden. Die Differenzen zwischen der Vorhersage und dem wahren Bild würden gesendet werden wie in einem herkömmlichen Bewegungskompensationssystem.

Das System sollte eine hohe Datenkompression ohne eine nennenswerte Bildverschlechterung aus mehreren Gründen erzielen:

1) Wenn ein Gegenstand verdeckt ist und dann wieder enthüllt wird, müssen die Daten nicht erneut übertragen werden.
2) Für sich relativ wenig ändernde Körper, wie z.B. Rümpfe, könnte eine sehr gute Vorhersage gebildet werden unter Verwendung von weniger umfangreichen graphischen Transformationen, z.B. Translationen und Rotationen in der Bildebene und Maßstabsänderungen. Die Differenzen zwischen der Vorhersage und dem Wahren sollten klein sein.
3) Für die sich schneller bewegenden Gegenstände sollte eine gute Vorhersage immer noch möglich sein, obwohl die Differenzen größer wären.
4) Es könnte subjektiv wichtige Eigenarten in der Szene unterschiedlich von weniger wichtigen Eigenarten behandeln. Zum Beispiel könnten Gesichter stärker gewichtet werden als sich schnell bewegende Arme.

Ein zweites Ausführungsbeispiel der Erfindung bezieht sich auf die Synthese eines sich bewegenden Bildes eines Lautsprechers, um synthetisierte Sprache zu begleiten. Zwei Typen von Sprachsynthese werden betrachtet:

a) Synthese eines begrenzten Wortschatzes, wobei digitalisierte Darstellungen vollständiger Worte gespeichert werden, und die Worte unter Handsteuerung, Computersteuerung oder eine andere Eingabe abgerufen und regeneriert werden. Die Art der Speicherung, ob PCM oder z.B. als Formatparameter, berührt die Bildsynthese nicht.

b) Allophonsynthese, wobei jedes Wort synthetisiert werden kann durch Verwendung von Coden, die auszusprechende Klänge darstellen; diese Code können direkt erzeugt werden vom Eingabetext (Text-zu-Sprache-Systeme).

In jedem Fall besteht die Gesichtssynthese aus zwei Stufen; einer Lernphase, welche der oben beschriebenen entspricht, und einer Synthesephase, in welcher die entsprechenden Mund-Codeworte erzeugt werden, um die synthetisierte Sprache zu begleiten.

Betrachtet man Option (a) zuerst, wird das Sprachvokabular gewöhnlich erzeugt, indem man die Äußerungen eines einheimischen Sprechers aufzeichnet, und es wird oft geschickt sein, das Gesicht des gleichen Sprechers zu verwenden. Wird ein anderes Gesicht gewünscht oder um eine Sichteinrichtung zu einem existierenden System hinzuzufügen, kann der Ersatzsprecher zusammen mit einem erneuten Abspielen des Sprachvokabulars sprechen. Für jeden Fall ist das Vorgehen das gleiche. Die Lernphase ist die gleiche wie die oben beschriebene, dahingehend, daß das System den benötigten Gesichtsdatenblock und die Mundnachschlagetabelle erfaßt. Es muß jedoch auch die Sequenz von Mundpositions-Codeworten entsprechend jedem Wort aufzeichnen und diese Sequenz in einer weiteren Tabelle speichern (die Mundcodetabelle). Es wird hier bemerkt, daß dieses Vorgehen nicht im Echtzeitverfahren ausgeführt werden muß und bietet daher die Möglichkeit, die Mundsequenzen für jedes Wort zu optimieren.

In der Synthesephase werden an den Synthetisator bereitgestellte Eingabecodes nicht nur dazu verwendet, Sprachdaten wiederzugewinnen und sie an eine Sprachregenerierungseinheit oder einen Synthetisator weiterzuleiten, sondern auch, um die Mundecodeworte wiederzugewinnen und diese synchron mit der Sprache zu einem Empfänger zu übertragen, welcher die sich bewegenden Bilder rekonstruiert, wie oben mit Bezug auf Fig. 1 beschrieben. Alternativ könnten die Empfängerfunktionen lokal durchgeführt werden für eine lokale Anzeige oder für eine Weiterübertragung eines Standard-Videosignals.

In dem Fall der Allophonsynthese (b) wird wiederum ein echtes Gesicht benötigt, und die zuvor beschriebene Lernphase wird durchgeführt, um die Gesichtsbild- und Mundbild-Tabelle zu erzeugen. Hier ist es jedoch notwendig, Mundpositionen mit individuellen Phonemen (d.h. Teilen von Worten) zu korrelieren, und daher muß der Besitzer des Gesichtes gleichzeitig mit deren Erzeugung durch den Sprachsynthetisator einen repräsentativen Textabschnitt äußern einschließlich zumindest eines Beispiels jedes Allophons, welches von dem Sprachsynthetisator hergestellt werden kann. Die erzeugten Codeworte werden dann in eine Mundnachschlagetabelle eingetragen, in welcher jeder Eintrag einem Allophon entspricht. Die meisten Eingaben werden aus mehr als einem Codewort bestehen. In einigen Fällen können die Mundpositionen entsprechend eines gegebenen Phonems in Abhängigkeit von den vorhergehenden oder folgenden Phonemen variieren, und dies kann auch berücksichtigt werden. Eine Wiedergewinnung der Sprach- und Videodaten findet auf eine ähnliche Weise statt wie die oben beschriebene für die "Ganzwort"-Synthese.

Man beachte, daß in dem Ausführungsbeispiel der "synthetischen Sprache" der Gesichtsdatenblock, die Mundbildtabelle und die Mundpositions-

Codeworte wie in dem oben beschriebenen Übertragungssystem zu einem entfernten Empfänger übertragen werden können zum Erzeugen eines sich bewegenden Bildes, aber unter gewissen Umständen, wie z.B. einer visuellen Anzeige zum Begleiten einer Computerausgabe synthetischer Sprache, kann die Anzeige lokal sein, und daher kann die "Empfänger"-Verarbeitung auf dem gleichen Gerät durchgeführt werden wie die Tabellen- und Codeworterzeugung. Alternativ kann das synthetisierte Bild lokal erzeugt werden und ein konventionelles Videosignal zu einem entfernten Monitor übertragen werden.

Die Frage der Synchronisation wird nun weiter betrachtet.

Eine typische Text-zu-Sprache-Synthese weist die folgenden Schritte auf:

(a) Umwandlung von einfacher Texteingabe zu phonetischer Darstellung.
(b) Umwandlung phonetischer zu niederphonetischer Darstellung.
(c) Umwandlung von niederphonetischen zu Formatparametern. Eine typische Parametereineuerungs- bzw. update-Periode wäre 10 ms.

Dieses Ausmaß an Verarbeitung involviert ein Maß an Verzögerung; weiterhin haben einige Konversionsstufen eine inhärente Verzögerung, da die Umwandlung vom Zusammenhang abhängt (z.B. wo der Klang eines bestimmten Buchstabens beeinflußt wird durch diejenigen, die ihm folgen). Daher involviert der Syntheseprozess, daß man Warteschlangen- und Zeitgabenotwendigkeiten sorgfältig in Betracht zieht, um sicherzustellen, daß die synthetisierten Lippenbewegungen mit der Sprache synchronisiert werden.

Wo (wie oben behandelt) die visuelle Synthese die Allophondarstellung für die Eingabedaten von dem Sprachsynthetisator verwendet und, falls

der Sprachsyntheseprozess von diesem Niveau abwärts vorhersagbare Verzögerungen involviert, kann eine passende Zeitgabe einfach dadurch sichergestellt werden, indem man entsprechende Verzögerungen in der visuellen Synthese einführt.

Ein alternativer Vorschlag besteht darin, Kennmarken in den Sprachdarstellungen einzufügen. Dies könnte die Option gestatten, daß man Mundpositionen in den Quellentext programmiert, anstatt von (oder zusätzlich zu) der Verwendung einer Nachschlagetabelle zum Erzeugen der Mundpositionen von den Allophonen. Bei beiden Arten könnten Kennmarken, welche die präzisen Augenblicke angeben, bei denen Mundpositionen wechseln, in den Sprachdarstellungen aufrechterhalten werden bis hinab (sagen wir) zu dem niederphonetischen Niveau. Der Sprachsynthetisator erzeugt eine Warteschlange niederphonetischer Codes, welche dann zu Formatparametern umgewandelt werden und zu der Formatsynthetisator-Hardware weitergeleitet werden; wenn die Codes von der Warteschlange "abgezogen" werden, wird jede Kennmarke, nachdem der ihr vorhergehende Text gesprochen worden ist, zu dem visuellen Synthetisator weitergeleitet, um den ensprechenden Mundpositionswechsel zu synchronisieren.

Ein drittes Ausführungsbeispiel der Erfindung betrifft die Erzeugung eines sich bewegenden Gesichts, um die Eingabe echter Sprache zu begleiten. Es wird wiederum ein Ersatzsprecher benötigt, um das Gesicht und die Lernphase bereitzustellen, denn die Erzeugung der Mundbildtabelle geschieht wie zuvor. Die Erzeugung der Mundcodetabelle hängt von der Einrichtung ab, welche verwendet wird, um die eingegebene Sprache zu analysieren; eine Option ist jedoch die Verwendung von Spektrumsanalyse, um Sequenzen spektraler Parameter zu erzeugen (eine wohlbekannte

Technik), wobei die Codetabelle dazu dient, diese Parameter und Mundbilder zu korrelieren.

Ein Gerät für eine derartige Sprachanalyse ist in Fig. 12 gezeigt. Jedes Vokalphonem hat ein unterschiedliches visuelles Erscheinungsbild. Das visuelle Korrelationselement des auditiven Phonems wird ein Visem genannt (K W Berger - Speechreading: Principles and Methods. Baltimore: National Educational Press, 1972, pages 73-107). Viele der Konsonanten haben jedoch das gleiche visuelle Erscheinungsbild, und die gebräuchlichste Klassifizierung von Konsonantenvisemen hat nur 12 Kategorien. Dies bedeutet, daß kein sichtbarer Fehler auftritt, falls das System Phoneme verwechselt, welche zu der gleichen Kategorie gehören. Da bei der Bildung von Konsonanten weniger akustische Energie erzeugt wird als bei der Bildung von Vokalen, wäre es für einen Spracherkenner schwieriger, zwischen Konsonanten zu unterscheiden. Daher ist die Abbildung Viel-zu-Eins von Konsonantenphonemen zu Konsonantenvisemen für dieses System dem Zufall überlassen.

Ein Verfahren zur Sprachanalyse würde eine Filterbank 10 mit 14-15 Kanälen verwenden, welche den gesamten Sprachbereich decken. Die akustische Energie in jedem Kanal wird integriert unter Verwendung eines verlustbehafteten Integrierers 11, und die Ausgabe 12 wird mit der Videodatenblockrate (alle 40 ms) abgetastet bzw. erfaßt. Von einem Subjekt wird verlangt, daß es während einer Übungssequenz einen vollständigen Satz von Phonemklängen ausspricht, und die Filterbank analysiert die Sprache. Individuelle Sprachklänge werden identifiziert, indem man einen Schwellwert für die Energie über jedem Satz von Proben einführt. Die Probenwerte werden in einem Satz von Speicherorten 13 gespeichert, welche mit dem entsprechenden Namen des Phonems markiert werden. Diese bilden einen Satz von Schablonen, welche daraufhin

verwendet werden, Phoneme in einem unbekannten Sprachsignal von der gleichen Person zu identifizieren. Dies wird getan unter Verwendung der Filterbank, um die unbekannte Sprache bei der gleichen Abtastrate zu analysieren. Die unbekannte Sprachprobe wird mit jeder der Schablonen verglichen, indem die Quadrate der Differenzen der entsprechenden Komponenten summiert werden. Die beste Übereinstimmung wird durch die kleinsten Differenzen gegeben. Damit gibt die Vorrichtung einen Code aus, der der besten Phonemübereinstimmung entspricht. Es würde auch einen speziellen Code geben, der Stille angibt.

Während das ubjekt den Satz Phoneme während der Übungssequenz äußerte, wird eine sich bewegende Sequenz von Bildern des Mundgebietes eingefangen. Durch Festlegen des Auftretens von jedem Phonem wird der entsprechende Datenblock in der Sequenz lokalisiert, und eine Teilmenge dieser Datenblöcke wird verwendet, um ein Codebuch von Mündern zu konstruieren. Im Betrieb wird eine Nachschlagetabelle verwendet, um den entsprechenden Mundcode durch den durch den Sprachanalysator erzeugten Code zu finden. Der die Stille bezeichnende Code sollte eine vollkommen geschlossene Mundposition beinhalten. Eine synthetische Sequenz wird erzeugt, indem man den entsprechenden Mund dem Gesicht bei einer Videorate überlagert.

Wie es der Fall bei der synthetisierten Sprache war, kann die Verarbeitung des "Empfängers" lokal oder aus der Ferne geschehen. In letzterem Fall wird als eine zusätzliche Modifikation vorgeschlagen, daß die bei dem Sender gespeicherte Mundbildtabelle eine größere Anzahl von Eintragungen haben kann als normal zu dem Empfänger gesendet wird. Dies würde die Tabelle in die Lage versetzen, Mundformen zu beinhalten, welche im allgemeinen nur selten auftreten, aber in gewissen Arten von Sprache häufig auftreten können: zum Beispiel Formen, die Klän-

gen entsprechen, welche nur bei gewissen regionalen Akzenten auftreten. Ein Erkennen der spektralen Parameter, welche einem solchen Klang entsprechen, würde dann den dynamischen Erneuerungsprozess, auf den zuvor verwiesen wurde, einleiten, um die relevanten Mundformen dem Empfänger zur Verfügung zu stellen.

Der Aufbau entsprechender Anzeige-(Empfänger)-Anordnungen für die obigen Vorschläge wird nun weitergehend betrachtet (siehe Fig. 13).

Ein Datenblockspeicher 100 wird bereitgestellt, in welchen während der Lernphase der empfangene ruhende Datenblock von einem Eingabedecodiergerät 101 eingeben wird, während der "Mund"-Speicher 102 die gespeicherte Anzahl (sagen wir 25) von Mundpositionen speichert. Eine Ausleselogik 103 liest wiederholt den Inhalt des Datenblockspeichers und fügt synchronisierende Impulse hinzu, um einen Videomonitor 104 zu speisen. In der Übertragungsphase werden empfangene Codeworte an die Steuereinheit 105 bereitgestellt, welche ein Überschreiben des relevanten Gebietes des Datenblockspeichers 101 mit den entsprechenden Mundspeicherdaten steuert. Es ist klar, daß dieses Überschreiben schnell sein muß, so daß es nicht für den Betrachter sichtbar ist. Diese Effekte könnten verringert werden, indem man das Gebiet zum Abspeichern der neuen Daten in kleine Blöcke aufteilt und es in einer zufälligen oder vorbestimmten nicht-sequentiellen Art und Weise überschreibt. Wenn alternativ die Datenblock-Speicherarchitektur Fenster oder Geisterbilder aufweist, dann könnten diese zuvor mit den neu gespeicherten Bildern geladen werden und hinein- und hinausgeschaltet werden, um die entsprechende Bewegung zu erzeugen. In einigen Fällen kann es möglich sein, den Prozess zu vereinfachen, indem man eine x-y Verschiebung der Fenster/Geisterbilder verwendet.

Patentansprüche

1. Gerät zum Kodieren eines sich bewegenden Bildes einschließlich eines menschlichen Gesichts (5), welches aufweist:

eine Einrichtung (1) zum Empfangen von Videoeingabedaten;

eine Einrichtung zur Datenausgabe, welche es gestattet, einen Datenblock des Bildes wiederherzustellen;

eine im Betrieb für jeden Datenblock des Bildes angeordnete Identifikationseinrichtung zum Identifizieren des Teiles der Eingabedaten, welche dem Mund (6) des dargestellten Gesichts entsprechen und

(a) um in einer ersten Betriebsphase die Munddatenteile jedes Datenblocks mit denen anderer Datenblöcke zu vergleichen, um einen repräsentativen Satz (Fig. 3) von Munddatenteilen auszuwählen, den repräsentativen Satz zu speichern und diesen Satz auszugeben;

(b) um in einer zweiten Phase die Munddatenteile jedes Datenblocks mit denen des gespeicherten Satzes zu vergleichen und zum Erzeugen eines auszugebenden Codeworts, welches anzeigt, welchem Element des Satzes die Munddatenteile dieses Datenblocks am meisten ähneln.

2. Gerät nach Anspruch 1, in welchem die Identifikationseinrichtung im Betrieb angeordnet ist, um als erstes denjenigen Teil eines Datenblocks von Eingabedaten zu identifizieren, der dem Mund des dargestellten Gesichts entspricht und zum Identifizieren des Mundteils von nachfolgenden Datenblöcken durch Autokorrelation mit Daten des einen Datenblocks.

3. Gerät nach Anspruch 1 oder 2, welches angeordnet ist, um im Betrieb während der ersten Phase einen ersten Munddatenteil zu speichern und dann für

die Munddatenteile jedes nachfolgenden Datenblocks ihn mit dem ersten und jedem anderen gespeicherten Munddatenteil zu vergleichen, und, falls das Ergebnis des Vergleiches einen Schwellenwert überschreitet, ihn zu speichern und auszugeben.

4. Gerät nach Anspruch 1, 2 oder 3, in welchem der Vergleich von Munddaten durch Subtraktion individueller Bildelementwerte und Summieren der absoluten Werte der Differenzen durchgeführt wird.

5. Gerät nach Anspruch 1, 2, 3 oder 4 einschließlich einer Einrichtung zum Erhalten der Koordinaten der Position des Gesichts innerhalb nachfolgender Datenblöcke des Bildes und Erzeugen kodierter Daten, welche diese Koordinaten darstellen.

6. Gerät nach einem der vorhergehenden Ansprüche, in welchem während der zweiten Phase in dem Falle, daß das Ergebnis des Vergleichs zwischen einem Munddatenteil und demjenigen des Satzes, welchem es am meisten ähnelt, eine vorbestimmte Schwelle überschreitet, dieser Datenteil ausgegeben und als ein Teil des Satzes gespeichert wird.

7. Gerät nach einem der vorhergehenden Ansprüche, weiterhin mit einer Identifikationseinrichtung, welche angeordnet ist, im Betrieb für jeden Datenblock des Bildes denjenigen Teil der Eingabedaten zu identifizieren, der den Augen des dargestellten Gesichts entspricht, und

(a) in der ersten Betriebsphase die Augendatenteile jedes Datenblocks mit denen anderer Datenblöcke zu vergleichen, um einen repräsentativen Satz von Augendatenteilen auszuwählen, diesen repräsentativen Satz zu speichern und den Satz auszugeben;

(b) in der zweiten Phase den Augendatenteilen jedes Datenblocks mit denen des gespeicherten Satzes zu vergleichen und ein Codewort zu erzeugen, welches angibt, welchem Element des Satzes der Augendatenteil dieses Datenblocks am meisten ähnelt.

8. Sprachsynthetisator, welcher eine Einrichtung zur Synthese eines sich bewegenden Bildes beinhaltet einschließlich eines menschlichen Gesichts, wobei der Synthetisator aufweist:

(a) eine Einrichtung zum Speichern und Ausgeben des Bildes eines Gesichts;

(b) eine Einrichtung zur Speicherung und Ausgabe eines Satzes von Munddatenblöcken (Fig. 3), deren jede dem Mundgebiet des Gesichts entsprechen und eine jeweilige unterschiedliche Mundform darstellen;

(c) eine Eingabe zum Empfangen von Codes, welche Worte oder Teile von zu sprechenden Worten identifizieren;

(d) eine Sprachsyntheseeinrichtung, welche auf den an der Eingabe empfangenen Code anspricht, um Worte oder dazu entsprechende Teile von Worten zu synthetisieren;

(e) eine Einrichtung, die eine Tabelle speichert, welche derartige Codes mit Codeworten in Beziehung setzt, welche die Munddatenblöcke oder Sequenzen derartiger Codeworte identifiziert; und

(f) eine Steuereinrichtung, welche auf die an der Eingabe empfangenen Codes anspricht, um das entsprechende Codewort oder die Codewortsequenz von der Tabelle auszuwählen und sie synchron mit der Synthese des entsprechenden Wortes oder Teiles eines Wortes von der Sprachsyntheseeinrichtung auszugeben.

9. Synthetisator nach Anspruch 8, in welchem die Sprachsyntheseeinrichtung eine Einrichtung beinhaltet, die angeordnet ist, um im Betrieb die Eingabecodes zu verarbeiten und in Warteschlangen einzureihen, wobei die Warteschlange Kennzeichencodes enthält, welche Änderungen in der Mundform anzeigen, und in Antwort auf jeden Kennzeichencode zum Senden einer Anzeige an die Steuereinrichtung, nachdem der Sprachsynthetisator die Sprache erzeugt hat, welche durch den Eingabecode dargestellt wird, der dem Kennzeichencode in

der Warteschlange vorausgeht, wobei die Steuereinrichtung das an die synthetisierte Sprache ausgegebene Codewort synchronisieren kann.

10. Gerät zur Synthese eines sich bewegenden Bildes, wobei das Gerät aufweist:

(a) eine Einrichtung zum Speichern und Ausgeben des Bildes eines Gesichts;

(b) eine Einrichtung zum Speichern und Ausgeben eines Satzes von Munddatenblöcken, die jeweils dem Mundgebiet des Gesichts entsprechen und eine jeweilige unterschiedliche Mundform darstellen;

(c) eine Audioeingabe zum Empfangen von Sprachsignalen und einer Frequenzanalyseeinrichtung (10, 11, 12), welche auf derartige Signale anspricht zum Erzeugen von Sequenzen spektraler Parameter;

(d) eine Einrichtung (13), die eine Tabelle speichert, welche spektrale Parametersequenzen mit Codeworten in Beziehung setzt, wobei Munddatenblöcke oder Sequenzen davon identifiziert werden;

(e) eine Steuereinrichtung, die auf die spektralen Parameter anspricht, um für eine Ausgabe die entsprechenden Codeworte oder Codewortsequenzen von der Tabelle auszuwählen.

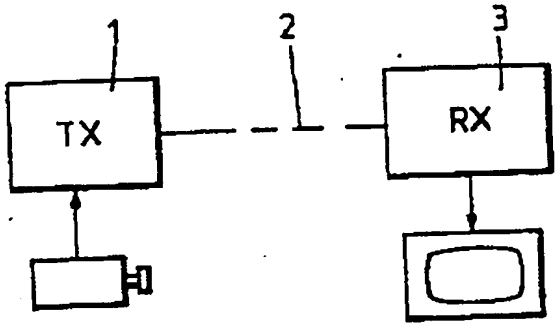
11. Gerät nach Anspruch 8, 9 oder 10, weiterhin mit einer Datenblockspeichereinrichtung (100) zum Empfangen und Speichern von Daten, welche einen Datenblock des Bildes darstellen;

eine Einrichtung (103) zum repetitiven Auslesen des Datenblockspeichers zum Erzeugen eines Videosignals; und

eine Steuereinrichtung (105), welche angeordnet ist, um im Betrieb die ausgewählten Codeworte zu empfangen und in Antwort auf jedes Codewort den entsprechenden Munddatenblock auszulesen und ein Einfügen dieser Daten in die Daten, welche der Leseeinrichtung (103) bereitgestellt werden, zu bewirken.

FÜR D. ANMELDER(IN):
12. FEB. 1998
PATENTANWÄLTE

Fig.1.



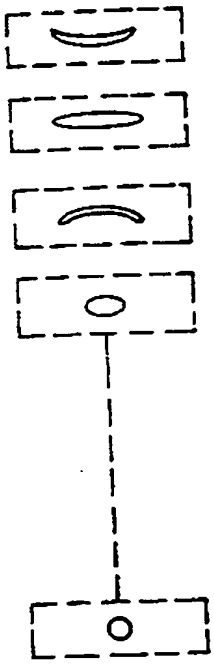


Fig.3.

Fig.2.

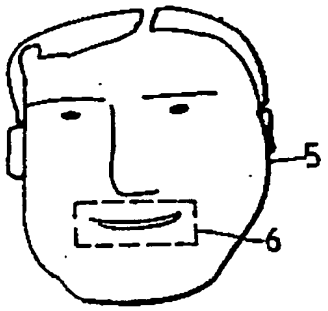


Fig.13.

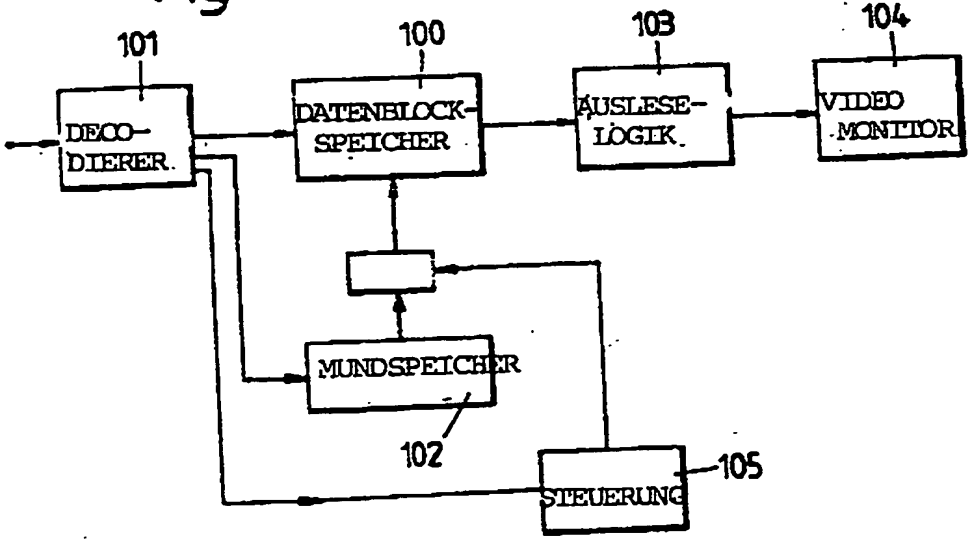


Fig.4.

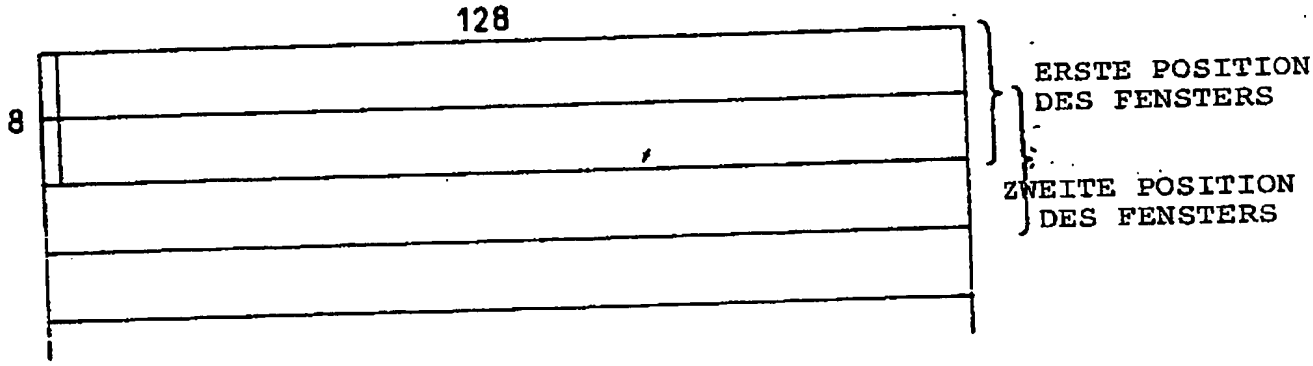

128
8
ERSTE POSITION
DES FENSTERS
ZWEITE POSITION
DES FENSTERS

Fig.5.

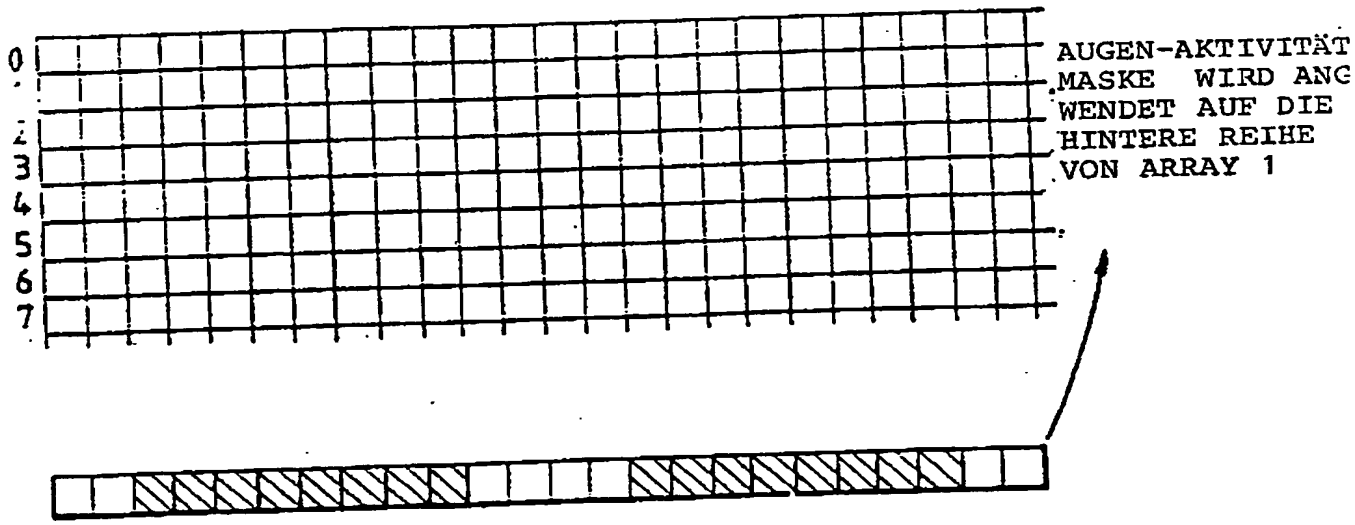

0
1
2
3
4
5
6
7
AUGEN-AKTIVITÄT
MASKE WIRD ANG
WENDET AUF DIE
HINTERE REIHE
VON ARRAY 1

Fig.6.

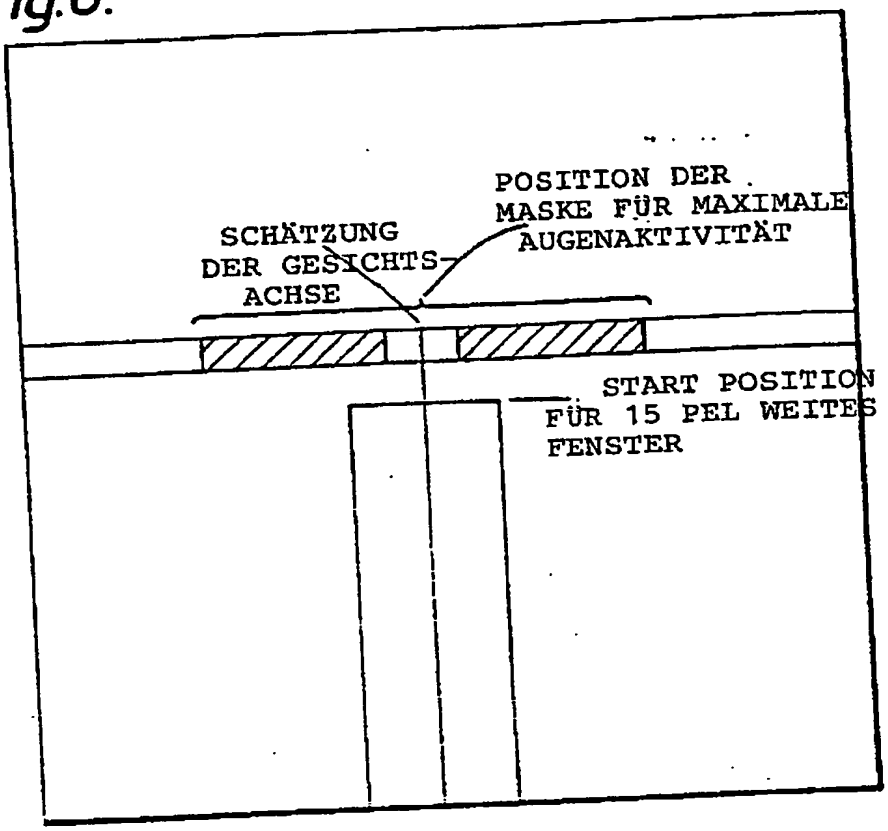
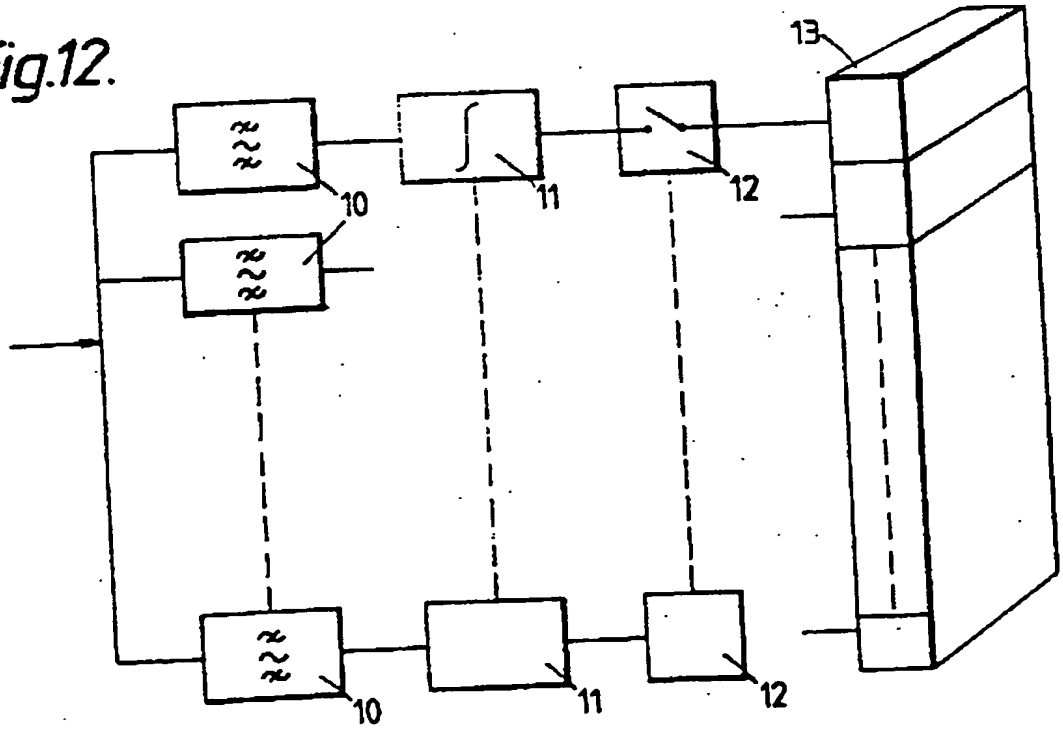

POSITION DER
MASKE FÜR MAXIMALE
AUGENAKTIVITÄT
SCHÄTZUNG
DER GESICHTS-
ACHSE
START POSITION
FÜR 15 PEL WEITES
FENSTER

Fig.12.


13
10
11
12
10
11
12

```
 0  000000
 1  000000
 2  00000
 3  0000
 4  0000
 5  0000
 6  000000
 7  00000
 8  00000
 9  00000
10  00000
11  00000
12  00000
13  0000
14  000
15  00
16  0
17  0
18  00000
19  000000000
20  00000000000
21  000000000000000
22  0000000000000000      <-------  UNTERER TEIL DER NASE
23  000000000000000
24  00000000000
25  0000000000
26  000000
27  000000
28  00000
29  000
30
31
32
33  0000
34  000000000000
35  0000000000000000      <-------  MUND
36  0000000000000000
37  000
38
39
40
41  0000
42  0000000000000000      <-------  UNTERLIPPE
43  000000000000000
44  0000000000
45  000000
46  0000
47  000
48
49  0
50  0
51  0
52  0
53
54
55
56
57
```

Fig. 7.

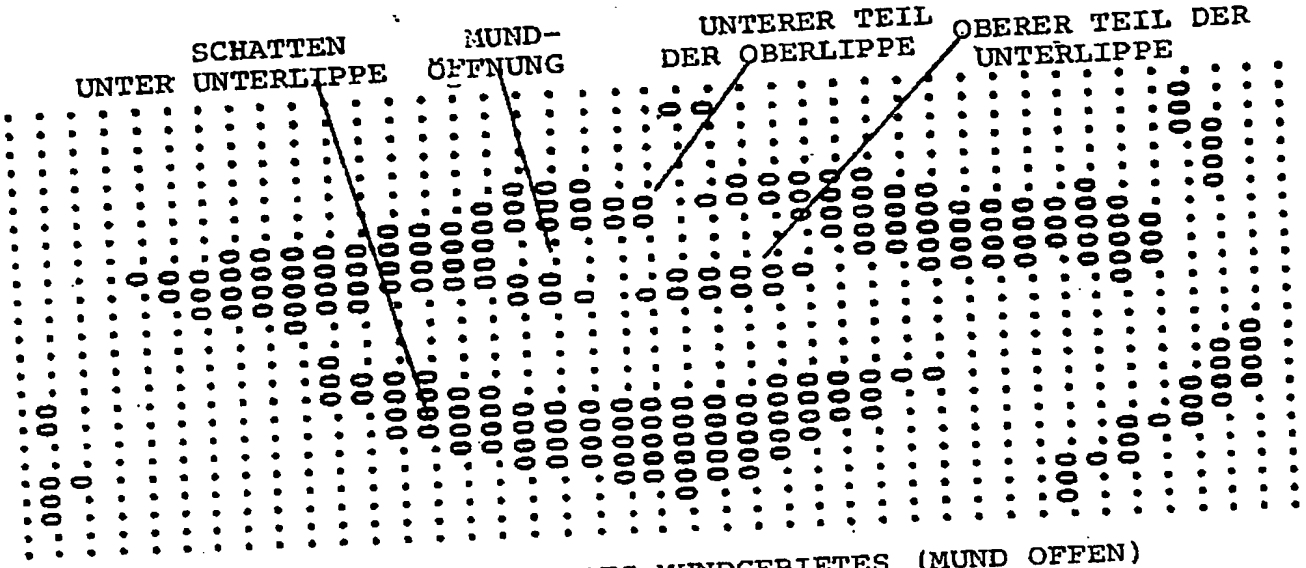


Fig.8. BINÄRES BILD DES MUNDGEBIETES (MUND OFFEN)
(o - SCHWARZE PELS, . WEISSE PELS)

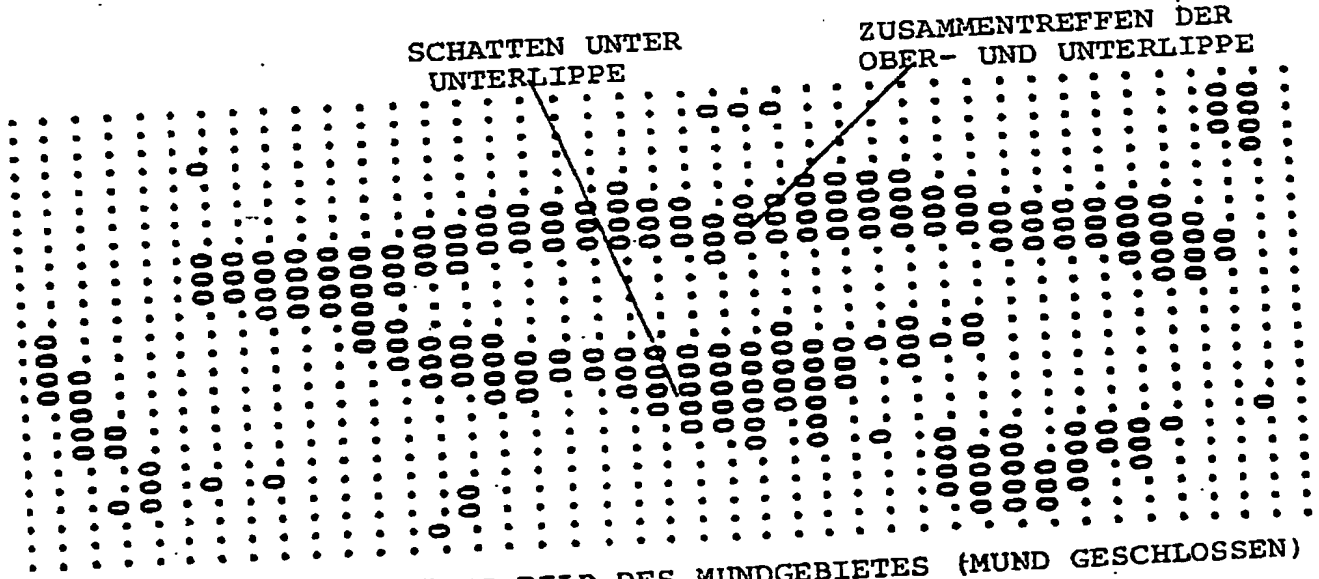


Fig.9. BINÄRES BILD DES MUNDGEBIETES (MUND GESCHLOSSEN)
o - SCHWARZE PELS, . WEISSE PELS)

Fig.10.

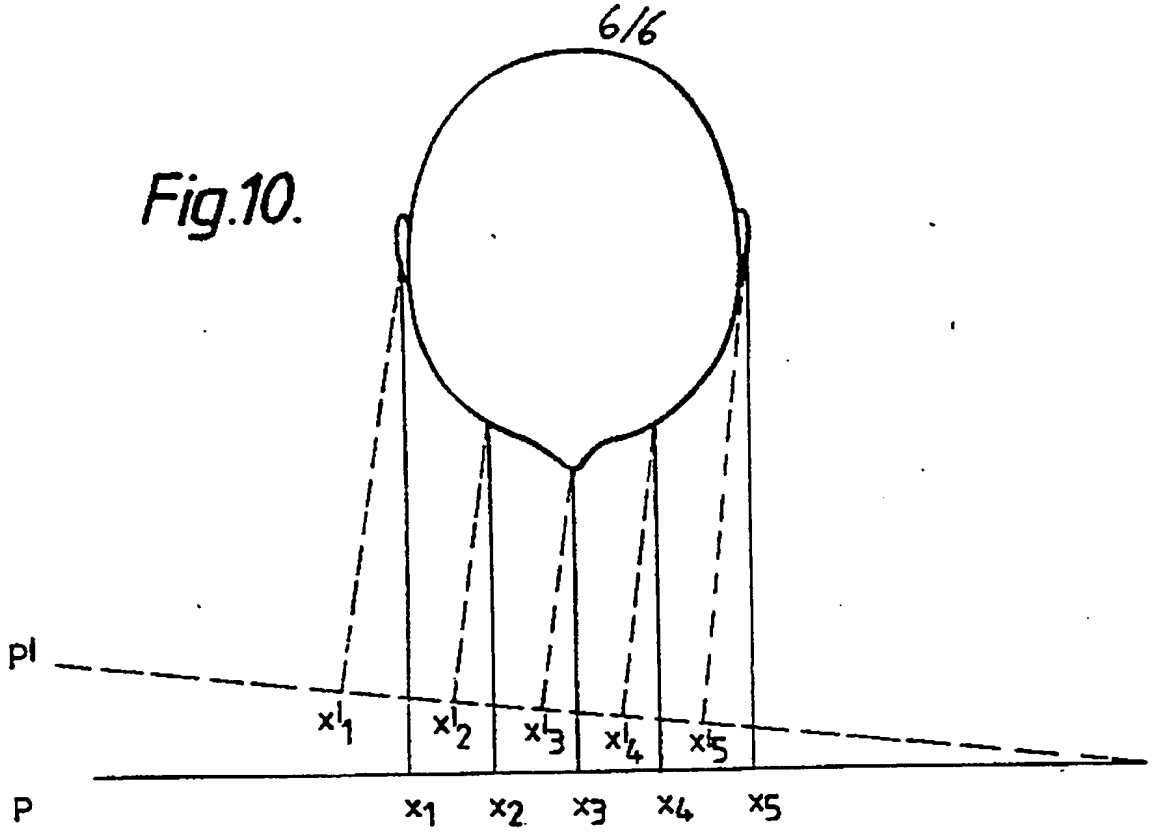
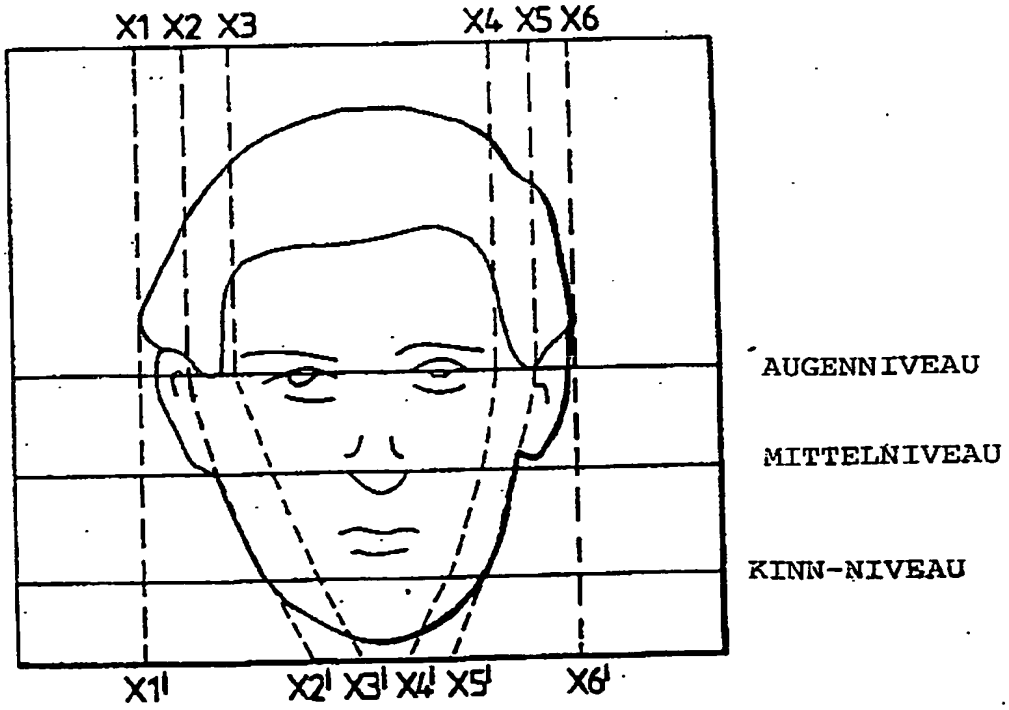

pl
P
x'1
x'2
x'3
x'4
x'5
x1
x2
x3
x4
x5

Fig.11.


X1 X2 X3
X4 X5 X6
AUGENNIVEAU
MITTELNIVEAU
KINN-NIVEAU
X1'
X2' X3' X4' X5'
X6'

This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☐ FADED TEXT OR DRAWING
- ☐ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.
As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.